Undergraduate Thesis Proposal

Implementation of Model to Measure Evolutionary Rate Variation In Proteins

Student - Caio Fontes de Castro
Advisor - Dr. Milton Yutaka Nishyama Jr.

April 30 of 2022

1 Proposal

Synonymous mutations in protein sequences are defined as substitutions in the genome level that either do not alter the aminoacid being expressed in the translation process, or alter it to an aminoacid with similar physicochemical properties [Muse and Gaut, 1994]. Given the rate at which random mutations occur in the genome, conservation in aminoacid sites inside proteins is indicative of a positive selection pressure acting in this site [Rubinstein and Pupko, 2012].

These synonymous mutations were often not given attention by biologists, since they do not alter the protein content or function [King and Jukes, 1969], and based on that assumption the mathematical models of protein evolution defined the synonymous mutation rate as a fixed constant dS for all sites [Muse and Gaut, 1994]. This assumption has since then been challenged, and models that ascribe the same importance to synonymous and non-synonymous rates of variation consistently ouperform models that don't [Wisotsky et al., 2020].

In this project we will develop an efficient implementation of a state of the art model for calculating the synonymous mutation rate accross protein sites.

2 Literature Review

In Evolutionary Biology, since it's not possible to go into the past and directly observe the changes happening in the molecular level, it's common to compare similar proteins among multiple species through a Multiple Sequence Alignment (MSA) of their aminoacid sequences, and then combine this information with the estimated phylogenetic tree of the samples to try and reconstruct the history of the given protein.

In these alignments we can see that not all regions of the protein are kept similar in multiple organisms, or "conserved". Some regions are highly conserved, and other have differing degrees of variability in their sequence. From these observations we can define the "evolutionary rate" of each protein site (aminoacid position) as the number of substitutions per site per year, and we can also observe significant variation in this quantity across proteins and across protein sites [Pupko et al., 2002].

Synonymous mutations are comprised by two kinds of events. In the first case, the mutation in the genome level (nucleotide change) does not change the expressed sequence of the protein (amino acid change). This is due the fact that multiple codons (triplets of nucleotides in DNA) can

generate the same aminoacid in the translation process due to some redundancy in the enconding process. The second kind of event is when the mutation in the genome changes the expressed aminoacid to another aminoacid that is similar in physicochemical characteristics, thus maintaining the function/structure of the protein.

For a long time, synonymous mutations were assumed to be "neutral", since they do not impact the function of the protein. However recent developments have shown that they are actually evidence of positive selection pressures, and can in fact have an influence in DNA/RNA interactions and in translational efficiency. [Bailey et al., 2021, Wisotsky et al., 2020, D'Andrea et al., 2019]

The mathematical models for estimation of these rates, for synonymous and nonsynonymous mutations are varied. The first models for estimating these quantities utilized a constant for the synonymous variation rate, considering that only non-synonymous mutations reflected evolutionary pressures [Muse and Gaut, 1994].

In [Pond and Muse, 2005] both rates are modelled as independent random variables that follow gamma distributions. [Mayrose et al., 2007] then developed a model to account for depedencies between adjacent sites, assumed to have independent variation rates previously. The model is based on two hidden Markov models that operate on the spatial dimension: one describes the dependency between adjacent non-synonymous rates while the other describes the dependency between adjacent synonymous rates. Finally, to account also for the dependencies between the non-synonymous ans synonymous variation rates in the same site the model in [Rubinstein and Pupko, 2012] was developed.

3 Workflow

In this project we plan to implement a command line program that efficiently performs the estimation of synonymous and non-synonymous rates of protein sites given the MSA and the associated phylogenetic tree.

We will also conduct a case study of Arachind RNA-seq samples, obtained from the NCBI database, build the phylogenetic tree using stablished techniques, perform the MSA, and search for novel conserved domains in the data as a prove of concept for how our software can integrate in a traditional bioinformatics workflow.

3.1 Schedule

The proposed schedule is :

Step	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Literature Review	X	X						
Implementation and Testing		X	X	X	X			
Case Study					X	X		
Writing Thesis							X	X

3.2 Main Steps Description

- **Literature Review**: Review of the state of the art and definition of which theoretical evolutionary rate variation model will be implemeted. Both theoretical and computational limitations will be taken into consideration.

- **Implementation and Testing**: Implementation of the model and benchmarking with existing curated datasets, such as used in the cited models.
- **Case Study** : Application of the model in Arachnid samples form NCBI as a case study of the utilization of the software in a bioinformatics workflow.
- **Writing Thesis**: Writing of the graduate thesis document. The document will be codeveloped with the rest of the project, but this period will be almost exclusively dedicated to this process.

References

[Bailey et al., 2021] Bailey, S. F., Morales, L. A. A., and Kassen, R. (2021). Effects of synonymous mutations beyond codon bias: The evidence for adaptive synonymous substitutions from microbial evolution experiments. *Genome Biology and Evolution*, 13(9).

[D'Andrea et al., 2019] D'Andrea, L., Pérez-Rodríguez, F.-J., de Castellarnau, M., Guix, S., Ribes, E., Quer, J., Gregori, J., Bosch, A., and Pintó, R. M. (2019). The critical role of codon composition on the translation efficiency robustness of the hepatitis a virus capsid. *Genome Biology and Evolution*, 11(9):2439–2456.

[King and Jukes, 1969] King, J. L. and Jukes, T. H. (1969). Non-darwinian evolution. *Science*, 164(3881):788–798.

[Mayrose et al., 2007] Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. (2007). Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, 23(13):i319–i327.

[Muse and Gaut, 1994] Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*.

[Pond and Muse, 2005] Pond, S. K. and Muse, S. V. (2005). Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22(12):2375–2385.

[Pupko et al., 2002] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(Suppl 1):S71–S77.

[Rubinstein and Pupko, 2012] Rubinstein, N. D. and Pupko, T. (2012). Detection and analysis of conservation at synonymous sites. In *Codon Evolution*, pages 218–228. Oxford University Press.

[Wisotsky et al., 2020] Wisotsky, S. R., Pond, S. L. K., Shank, S. D., and Muse, S. V. (2020). Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: Ignore at your own peril. *Molecular Biology and Evolution*, 37(8):2430–2439.